Data Analysis and Manifold Learning Lecture 2: Properties of Symmetric Matrices and Examples

Radu Horaud
INRIA Grenoble Rhone-Alpes, France
Radu.Horaud@inrialpes.fr
http://perception.inrialpes.fr/

Outline of Lecture 2

- Basic definitions, eigen decomposition, LU and Cholesky matrix factorizations;
- Spectral decomposition, powers, inverse, exponential;
- Geometric interpretation;
- The Raleigh-Ritz theorem and extensions;
- Computing eigenvalues and eigenvectors in practice: power method, inverse power method, and shifted inverse power method;

Material for This Lecture

- R. A. Horn and C. R. Johnson. Matrix Analysis. Chapter 4: Hermitian and symmetric matrices.
- G. H. Golub and C. F. Van Loan. Matrix Computations. Chapter 8: The symmetric eigenvalue problem. Chapter 9: Lanczos methods.
- Software: `http://www.caam.rice.edu/software/ARPACK/` written in Fortran77!

Some Basic Definitions

- Symmetry of a $D \times D$ matrix: $\mathbf{A} = \mathbf{A}^\top$
- Eigen decomposition: $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with the properties:
  - $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_D$
  - $\det(\mathbf{U}) = \pm 1$
  - All the eigenvalues are real numbers:

$$\lambda_{\min} = \lambda_1 \leq \ldots \leq \lambda_i \leq \ldots \leq \lambda_D = \lambda_{\max}$$

- $\mathbf{A}$ is referred to as a *real symmetric matrix*;
- If $\lambda_1 \geq 0$ then it is a *positive semi-definite symmetric matrix*
- If $\lambda_1 > 0$ then it is a *positive definite symmetric matrix*
- Symmetric matrices are *nondefective*: the algebraic multiplicity of each eigenvalue is equal to its geometric multiplicity.

Spectral Decomposition, Deflation, Powers, Exponential

- A symmetric matrix can be written as $\mathbf{A} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ where $\{\boldsymbol{u}_i\}_{i=1}^{i=D}$ are the eigenvectors or, equivalently, the column vectors of $\mathbf{U}$.
- The transformation $\widetilde{\mathbf{A}} = \mathbf{A} - \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$ is known as a deflation.
- Note that $\widetilde{\mathbf{A}} \boldsymbol{u}_k = \mathbf{0}$.
- $\mathbf{A}^2 = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^\top$
- More generally: $\mathbf{A}^k = \mathbf{U}\boldsymbol{\Lambda}^k\mathbf{U}^\top$
- The matrices $\mathbf{A}, \mathbf{A}^2, \dots, \mathbf{A}^k$ have the same eigenvectors $\{\boldsymbol{u}_i\}$ and eigenvalues $\lambda, \lambda^2, \dots, \lambda^k$.
- Matrix exponential: $e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}$
- We have: $e^{\mathbf{A}} = \mathbf{U}\text{Diag}[e^{\lambda_1} \dots e^{\lambda_i} \dots e^{\lambda_D}]\mathbf{U}^\top$
- Hence, matrix $e^{\mathbf{A}}$ has the same eigenvectors as $\mathbf{A}$ and eigenvalues e^{λ}.

The Inverse of a Symmetric Matrix

- The case of a non-singular symmetric matrix:
  - $\mathbf{A}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top$.
  - Spectral decomposition: $\mathbf{A}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^\top$
  - The matrices $\mathbf{A}^{-1}, \mathbf{A}^{-2}, \ldots, \mathbf{A}^{-k}$ have eigenvectors $\boldsymbol{u}_i$ and eigenvalues $\lambda_i^{-1}, \lambda_i^{-2}, \ldots, \lambda_i^{-k}$

The Pseudoinverse of a Singular Symmetric Matrix

- If a symmetric matrix is singular, it has as eigenvalue $\lambda = 0$ with multiplicity $m > 0$.
- We rearrange the eigenvalue-eigenvector pairs and we retain the non-zero pairs, such that:

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}^\top$$

  with: $\tilde{\mathbf{\Lambda}} = \text{Diag}[\lambda_1 \dots \lambda_{D-m}]$ and $\tilde{\mathbf{U}} = [\boldsymbol{u}_1 \dots \boldsymbol{u}_{D-m}]$.
- $\tilde{\mathbf{U}}$ is a $D \times (D-m)$ matrix whose columns are orthogonal. Notice that $\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}_{D-m}$ but $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \neq \mathbf{I}_D$!
- The **Moore-Penrose pseudoinverse** :

$$\mathbf{A}^\dagger = \tilde{\mathbf{U}}\text{Diag}[\lambda_1^{-1} \dots \lambda_{D-m}^{-1}]\tilde{\mathbf{U}}^\top$$

The Choleski Factorization

- We consider the case of positive **definite** symmetric matrices. They can be written as $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ but the choice of $\mathbf{B}$ is not unique.
- Any such matrix can be decomposed as: $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$ with $\mathbf{L}$ being a low-triangular matrix with nonnegative diagonal entries. This decomposition is unique.
- Complexity of Choleski decomposition algorithms for a $D \times D$ non singular matrix: D^3 FLOPS. This is twice more efficient than the LU decomposition.
- Let $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$. No matrix inversion needed to solve it! This can be rewritten as:

$$\begin{cases} \mathbf{L}\boldsymbol{y} = \boldsymbol{b} \\ \mathbf{L}^\top \boldsymbol{x} = \boldsymbol{y} \end{cases}$$

Matrix Norms

- The Frobenius norm:

$$\|\mathbf{A}\|_F^2 = \mathsf{tr}(\mathbf{A}^\top \mathbf{A}) = \mathsf{tr}(\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top) = \mathsf{tr}(\mathbf{\Lambda}^2) = \sum_{i=1}^{D} \lambda_i^2$$

- The spectral norm:

$$\max_{\boldsymbol{v}} \frac{\|\mathbf{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \left(\max_{\boldsymbol{v}} \frac{\boldsymbol{v}^\top \mathbf{A}^\top \mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{v}} \right)^{1/2} = \lambda_{\max}$$

(see the Raylegh-Ritz theorem below)

Geometric Interpretation

- Consider a positive definite symmetric matrix; In this case all the eigenvalues are strictly positive.
- Quadratic form for any vector $\boldsymbol{x} \neq 0$:

$$\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} = (\mathbf{U}^\top \boldsymbol{x})^\top \boldsymbol{\Lambda}(\mathbf{U}^\top \boldsymbol{x}) = \sum_{i=1}^{D} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x})^2$$

- Let's transform the data into another coordinate frame: $\boldsymbol{z} = \mathbf{U}^\top \boldsymbol{x}$; we obtain: $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} = \boldsymbol{z}^\top \boldsymbol{\Lambda}\boldsymbol{z}$.

$$\boldsymbol{z}^\top \boldsymbol{\Lambda}\boldsymbol{z} = (z_1/\lambda_1^{-1/2})^2 + \ldots (z_D/\lambda_D^{-1/2})^2 = C$$

- This is an ellipsoid with axes $\boldsymbol{u}_1 \ldots \boldsymbol{u}_D$ and with half eccentricities $\lambda_1^{-1/2} \ldots \lambda_D^{-1/2}$ (Remember PCA...)

The Raylegh-Ritz Theorem

Theorem

(Raylegh-Ritz). Let $\mathbf{A}$ *be a symmetric matrix with ordered eigenvalues, then:*

$$\lambda_1 \boldsymbol{x}^\top \boldsymbol{x} \leq \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} \leq \lambda_D \boldsymbol{x}^\top \boldsymbol{x} \;\; \forall \boldsymbol{x}$$

$$\lambda_{\max} = \lambda_D = \max_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \max_{\boldsymbol{x}^\top \boldsymbol{x} = 1} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$$

$$\lambda_{\min} = \lambda_1 = \min_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \min_{\boldsymbol{x}^\top \boldsymbol{x} = 1} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$$

Proof of the Raylegh-Ritz Theorem

- From the eigendecomposition: $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} = \sum_{i=1}^{D} \lambda_i \left((\mathbf{U}^\top \boldsymbol{x})_i\right)^2$
- Notice that: $\sum_{i=1}^{D} \left((\mathbf{U}^\top \boldsymbol{x})_i\right)^2 = \|\mathbf{U}^\top \boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x}$
- Using the fact that the eigenvalues can be ordered, we get the first part of the theorm.
- By dividing we obtain: $\lambda_{\min} \leq \frac{\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \leq \lambda_{\max}, \ (\boldsymbol{x} \neq 0)$
- with equalities when $\boldsymbol{x}$ is a λ_1 or λ_D eigenvector.
- We have: $\frac{\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = (\boldsymbol{x}^\top / \sqrt(\boldsymbol{x}^\top \boldsymbol{x}))\mathbf{A}(\boldsymbol{x}/\sqrt(\boldsymbol{x}^\top \boldsymbol{x}))$ and hence the minimization/maximization of the Raleigh quotient is equivalent to:

$$\begin{cases} \max_{\boldsymbol{x}} \boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} \\ \boldsymbol{x}^\top \boldsymbol{x} = 1 \end{cases}$$

What About the Remaining Eigenvalues and Eigenvectors?

- Let's restrict $\boldsymbol{x}$ to be orthogonal to the smallest eigenvector $\mathbf{u}_1$, i.e, $\mathbf{u}_1^\top \boldsymbol{x} = 0$:
- $\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = \sum_{i=2}^{D} \lambda_i \left((\mathbf{U}^\top \boldsymbol{x})_i \right)^2 \geq \lambda_2 \boldsymbol{x}^\top \boldsymbol{x}$
- with equality when $\boldsymbol{x} = \boldsymbol{u}_2$
- Therefore we obtain:

$$\lambda_2 = \min_{\substack{\boldsymbol{x}^\top \boldsymbol{x} = 1 \\ \boldsymbol{x}^\top \boldsymbol{u}_1 = 0}} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$$

$$\lambda_{D-1} = \max_{\substack{\boldsymbol{x}^\top \boldsymbol{x} = 1 \\ \boldsymbol{x}^\top \boldsymbol{u}_D = 0}} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$$

Computing Eigenvalues and Eigenvectors in Practice

- The *power method* estimates the largest eigenvalue/eigenvector pair or an *eigenpair*.
- The *power method + deflation* estimates the second largest eigenpair, etc.
- The *inverse power method* estimates the smallest eigenpair.
- The *shifted inverse power method* allows to obtain intermediate eigenpairs.
- The Lanczos method is an adaptation of the power method. It is very useful for large and sparse matrices. It is used by the ARPACK package.

The Power Method

- Input: A symmetric matrix $\mathbf{A}$ and a random vector $\boldsymbol{x}_0$.
- At each iteration k:
  1. Normalize $\boldsymbol{y}_k = \frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|}$ and
  2. $\boldsymbol{x}_{k+1} = \mathbf{A}\boldsymbol{y}_k$.
- Check for convergence: $\|\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\| < \varepsilon$
- Output: $\mathbf{u}_D = \boldsymbol{y}_{k+1}$ and $\lambda_D = \boldsymbol{y}_{k+1}^\top \mathbf{A}\boldsymbol{y}_{k+1}$

Justification of the Power Method

- Let $\boldsymbol{x}_0 = \sum_{i=1}^{D} \alpha_i \boldsymbol{u}_i$ hence we obtain after the first iteration: $\boldsymbol{x}_1 = \mathbf{A}\boldsymbol{x}_0 = \sum_{i=1}^{D} \alpha_i \lambda_i \boldsymbol{u}_i$
- Normalize this vector: $\boldsymbol{y}_1 = \frac{1}{\beta_1} \sum_{i=1}^{D} \alpha_i \lambda_i \boldsymbol{u}_i$
- More generally: $\boldsymbol{y}_{k+1} = \frac{1}{\beta_1 \dots \beta_{k+1}} \sum_{i=1}^{D} \alpha_i \lambda_i^{k+1} \boldsymbol{u}_i$
- At the limit this vector becomes the "largest" eigenvector:

$$\begin{aligned}
\boldsymbol{y}_\infty &= \lim_{k\to\infty} \frac{\alpha_D \lambda_D^{k+1}}{\beta_1 \dots \beta_{k+1}} \left(\sum_{i=1}^{D-1} \frac{\alpha_i}{\alpha_D} \frac{\lambda_i^{k+1}}{\lambda_D^{k+1}} \boldsymbol{u}_i + \boldsymbol{u}_D \right) = \boldsymbol{u}_D \\
\lambda_D &= \boldsymbol{y}_\infty^\top \mathbf{A} \boldsymbol{y}_\infty
\end{aligned}$$

The Power Method with Deflation

- Consider the matrix $\widetilde{\mathbf{A}} = \mathbf{A} - \lambda_D \boldsymbol{u}_D \boldsymbol{u}_D^\top$
- Notice that $(0, \boldsymbol{u}_D)$ is an eigenpair of $\widetilde{\mathbf{A}}$ and that the remaining eigenpairs remain unchanged (refer to the spectral decomposition of $\mathbf{A}$ and to the fact that eigenvectors corresponding to **distinct** eigenvalues are orthogonal).
- It follows that the second largest eigenpair $(\lambda_{D-1}, \boldsymbol{u}_{D-1})$ of $\mathbf{A}$ becomes the largest eigenpair of $\widetilde{\mathbf{A}}$
- The power method can now be applied to $\widetilde{\mathbf{A}}$, etc.

The Inverse Power Method

- The smallest eigenvector-eigenvalue pair $(\mathbf{u}_1, \lambda_1)$ of $\mathbf{A}$ corresponds to the largest eigenvector-eigenvalue pair $(\mathbf{u}_1, \lambda_1^{-1})$ of $\mathbf{A}^{-1}$.
- The k-th iteration of the power method becomes: $\boldsymbol{x}_{k+1} = \mathbf{A}^{-1}\boldsymbol{y}_k$
- which can be written as: $\mathbf{A}\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$
- This can be solved using the Choleski factorization $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$:

$$\begin{cases} \mathbf{L}\boldsymbol{z} = \boldsymbol{y}_k \\ \mathbf{L}^\top \boldsymbol{x}_{k+1} = \boldsymbol{z} \end{cases}$$

The Shifted Inverse Power Method

- Let's consider the matrix $\mathbf{B} = \mathbf{A} - \alpha\mathbf{I}$ as well as an eigenpair $\mathbf{A}\boldsymbol{u} = \lambda\boldsymbol{u}$.
- $(\lambda - \alpha, \boldsymbol{u})$ becomes an eigenpair of $\mathbf{B}$, indeed:

$$\mathbf{B}\boldsymbol{u} = (\mathbf{A} - \alpha\mathbf{I})\boldsymbol{u} = (\lambda - \alpha)\boldsymbol{u}$$

  and hence $\mathbf{B}$ is a **real symmetric** matrix with eigenpairs $(\lambda_1 - \alpha, \boldsymbol{u}_1), \dots (\lambda_i - \alpha, \boldsymbol{u}_i), \dots (\lambda_D - \alpha, \boldsymbol{u}_D)$
- If $\alpha > 0$ is choosen such that $|\lambda_j - \alpha| \ll |\lambda_i - \alpha| \; \forall i \neq j$ then $\lambda_j - \alpha$ becomes the smallest (in magnitude) eivenvalue.
- The inverse power method (in conjuction with the *LU decomposition* of $\mathbf{B}$) can be used to estimate the eigenpair $(\lambda_j - \alpha, \boldsymbol{u}_j)$.